Predicting Weather at RDU
using Time Series Modeling

Danny Ross
Andrew Jin
Diya Mirji

Agenda

01 Data

02 Data Analysis

03 Feature Engineering

04 Linear Regression

05 Random Forest

06 Model Performance

Data

Hourly Temperature from RDU

- Hourly weather data from RDU Airport
    - Meteostat’s bulk data
- Temperature, Humidity, Precipitation, etc.
- Only kept temperature and year-month-day-hour -> Datetime index

Data from past years

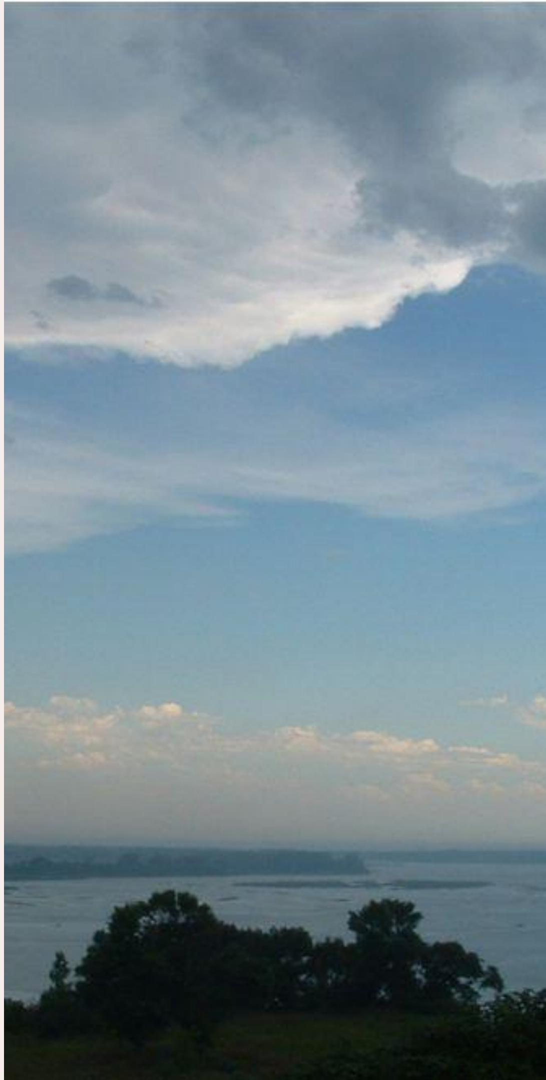
- Pulled all data from January 1, 2020 to September 30, 2025
    - Similar trend in temperature for each year

Train and Test Split

- Test Dataset: September 17, 2025 - September 30, 2025
- Train Dataset: January 1, 2020 - September 16, 2025

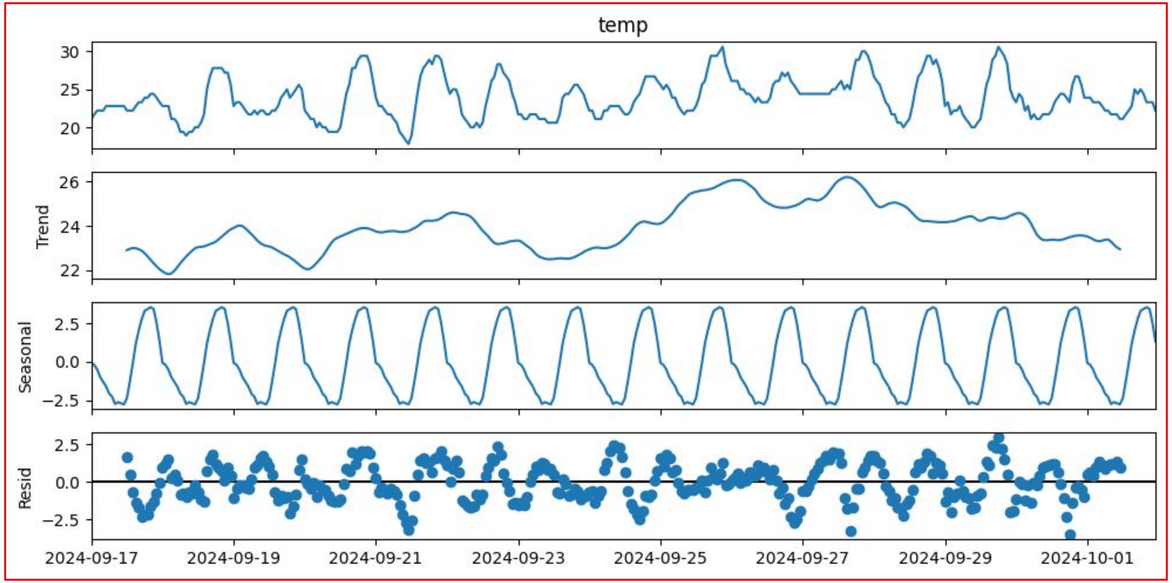
Training data size: **50064 samples**

Testing data size: **336 samples**



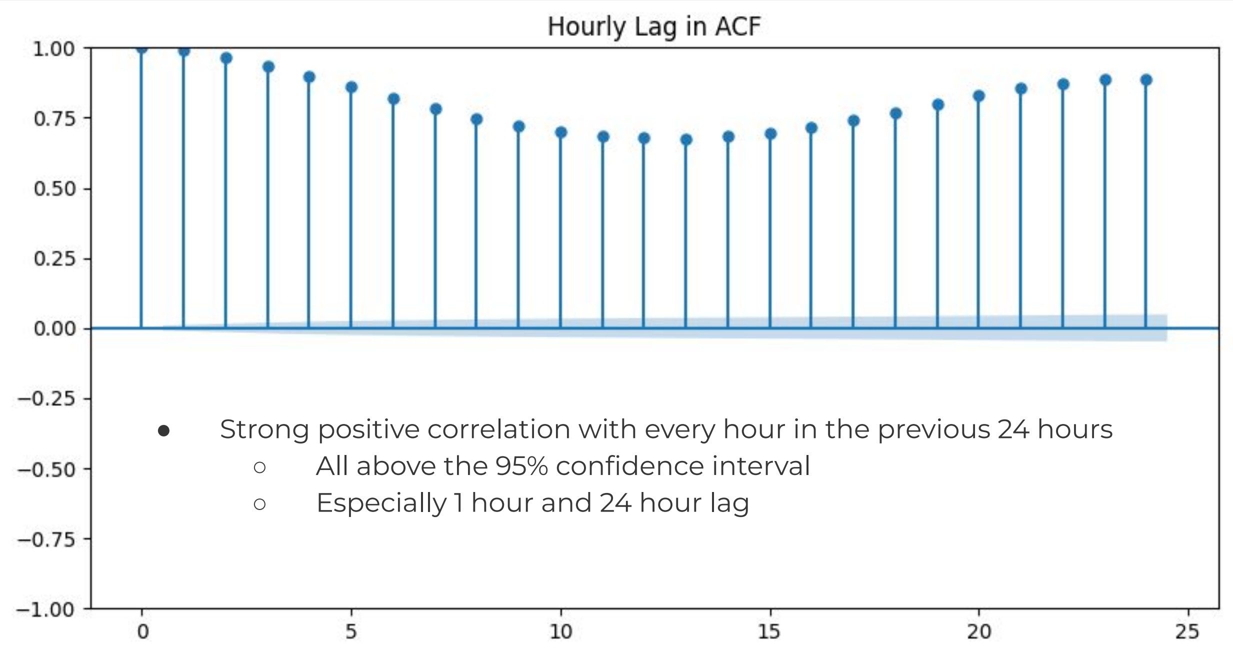
Data Analysis

- No missing data
- Roughly symmetric distribution with slight left skew
- Seasonal Decomposition of temperature during September 17, 2024 - September 30, 2024
    - Non linear trend and daily seasonality



Data Analysis

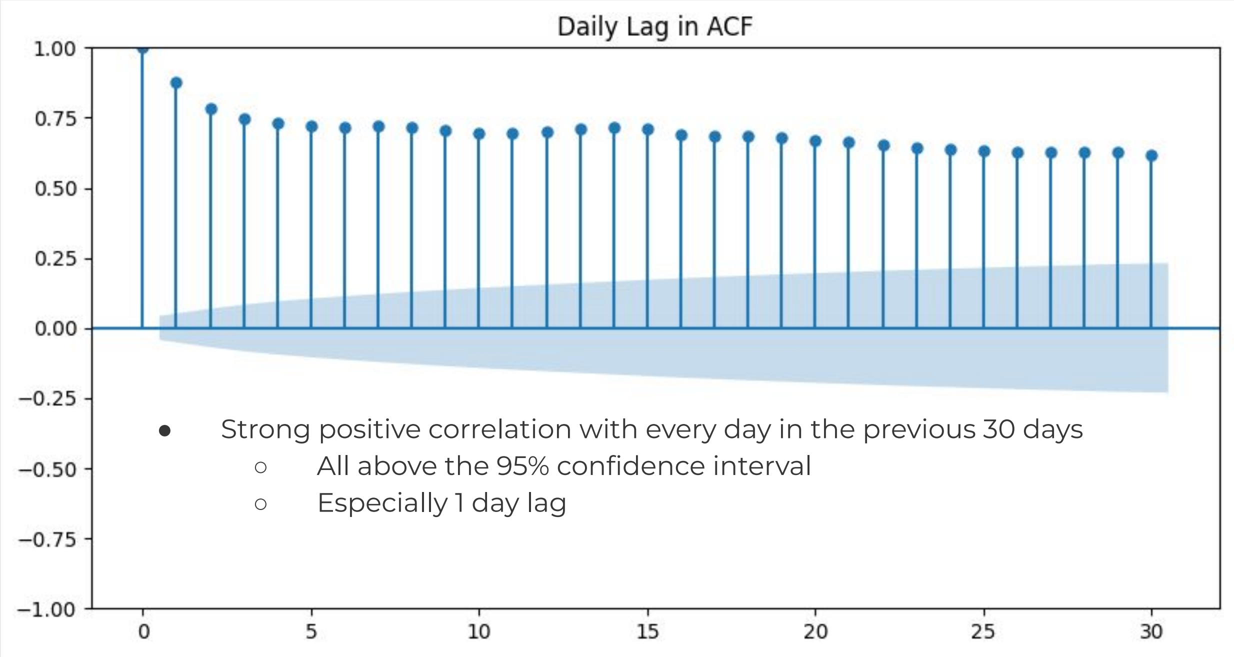
Autocorrelation



- Strong positive correlation with every hour in the previous 24 hours
    - All above the 95% confidence interval
    - Especially 1 hour and 24 hour lag

Data Analysis

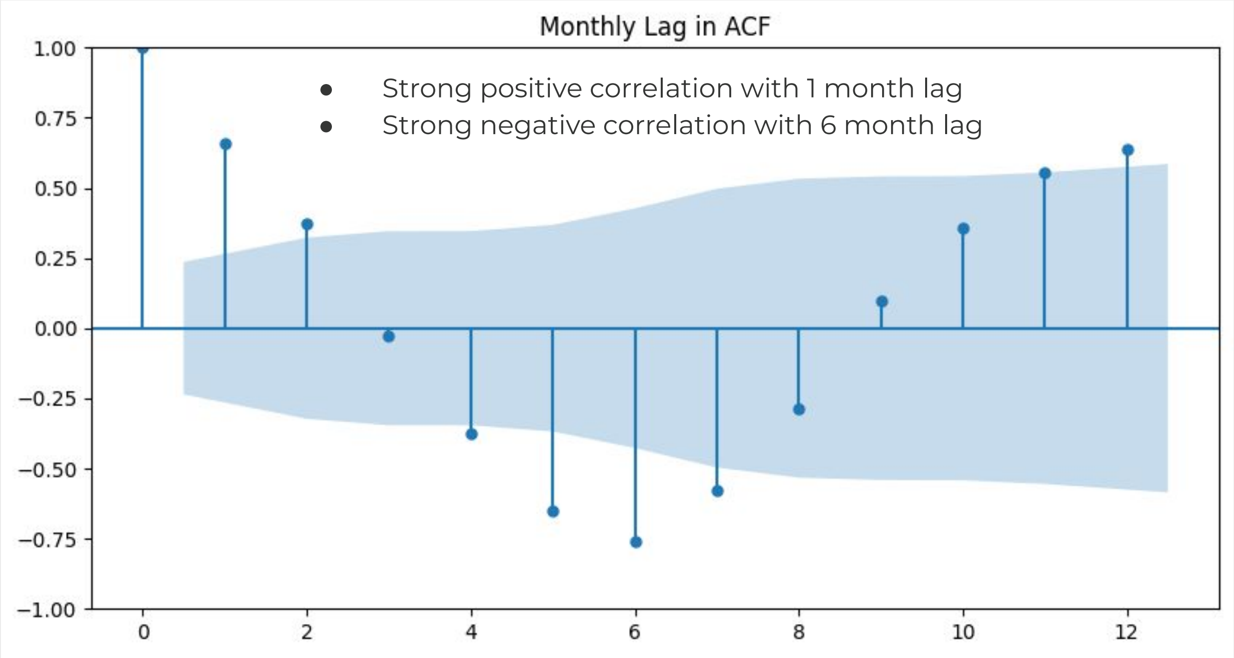
Autocorrelation



- Strong positive correlation with every day in the previous 30 days
  - All above the 95% confidence interval
  - Especially 1 day lag

Data Analysis

Autocorrelation

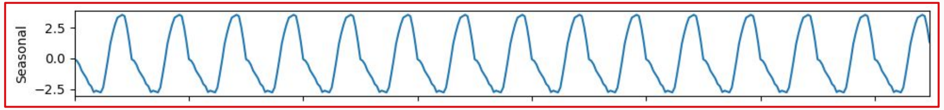


Feature Engineering

Pure Calendar Features

- Year (2020-2025)
- Month (1-12)
- Day (1-31)
- Hour (0-23)
- Day of Year (1-336)
- Day of Week (0-6)

Sine / Cosine Transformation



Cyclical Calendar Features

- captures seasonal cycle

- recognizes consecutive seasons

- hour_sin, hour_cos
  - smooth cycle of daily temperature
  - even between hour 23 and hour 0
- doy_sin, doy_cos
  - smooth cycle of temperature over year
  - even between day 365 and day 1
- month_sin, month_cos
  - smooth cycle of monthly temperature
  - even between month 12 and month 1

Harmonic Calendar Features

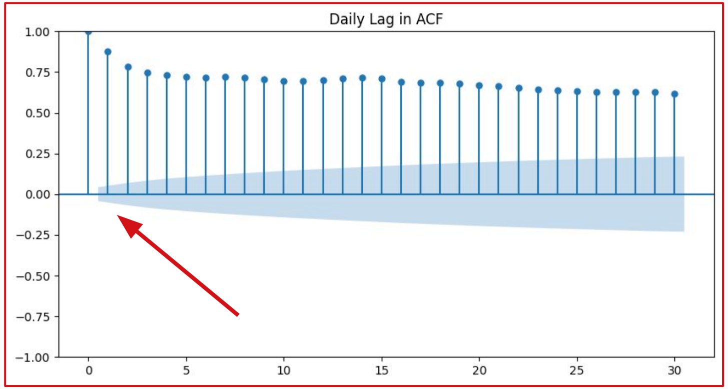
- captures multiple peaks per cycle

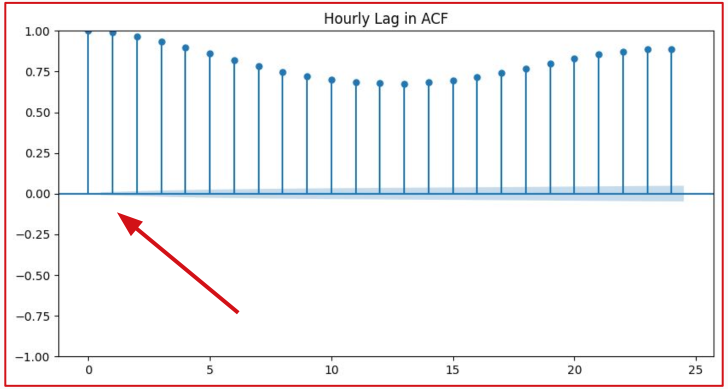
- hour_sin_2, hour_cos_2
  - 2x-daily peaks pattern
- hour_sin_3, hour_cos_3
  - 3 peaks in a day pattern
- doy_sin_2, doy_cos_2
  - 2-annual pattern
- doy_sin_3, doy_cos_3
  - 3 peaks per year pattern

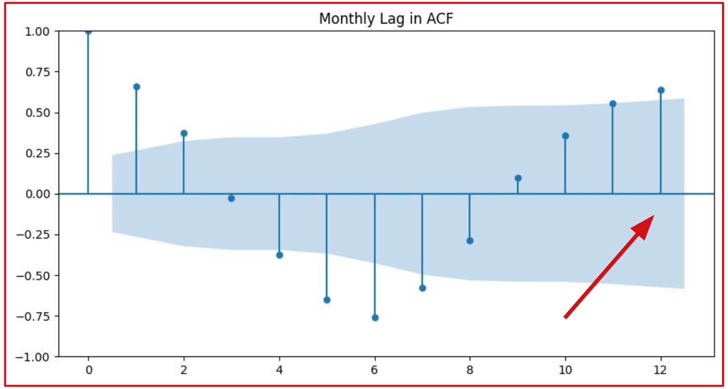
Feature Engineering

Autocorrelation Lags

- previous_hour_temp
  - previous hour's temperature
- previous_day_temp
  - previous day's temperature at same hour
- previous_year_temp
  - previous year's temperature at same month/day/hour

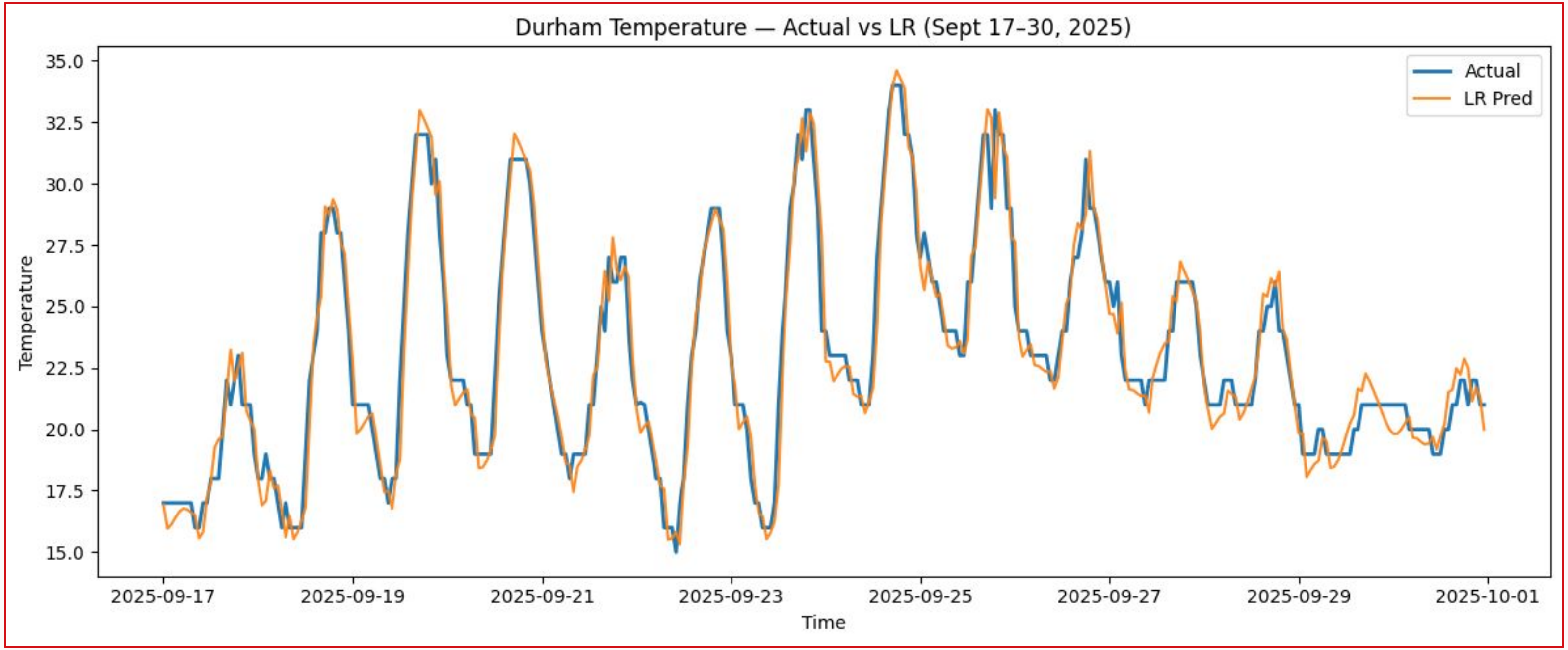




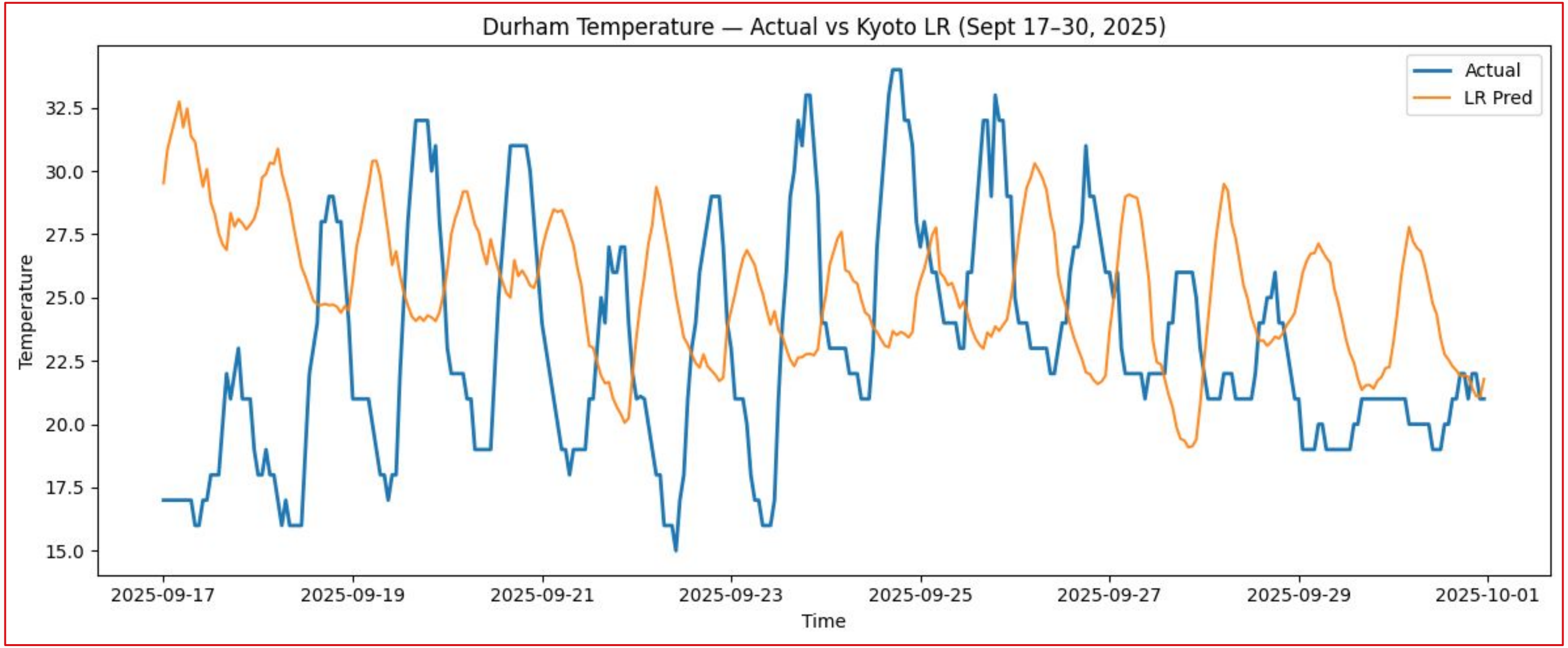


Linear Regression Model

- Most important features: the temperature from the **previous hour** (w=0.972), the **cosine of the hour** (w=-0.821), and the **sine of the hour** (w=-0.791).



Kyoto Model

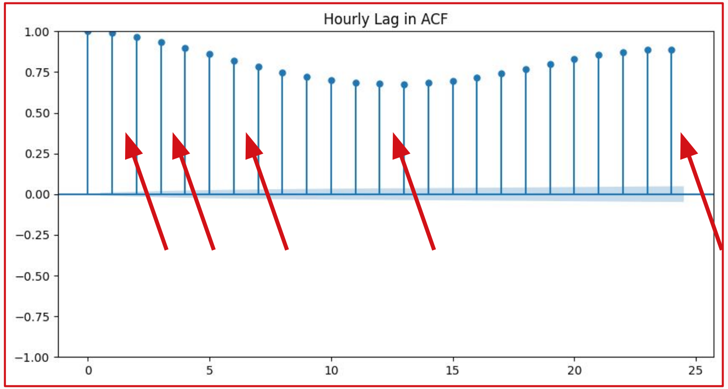


Durham Temperature — Actual vs Kyoto LR (Sept 17–30, 2025)
Actual
LR Pred
Temperature
32.5
30.0
27.5
25.0
22.5
20.0
17.5
15.0
2025-09-17
2025-09-19
2025-09-21
2025-09-23
2025-09-25
2025-09-27
2025-09-29
2025-10-01
Time

Feature Engineering pt2

More Lag Features

- Lag at hours 1, 3, 6, 12, 24, 48

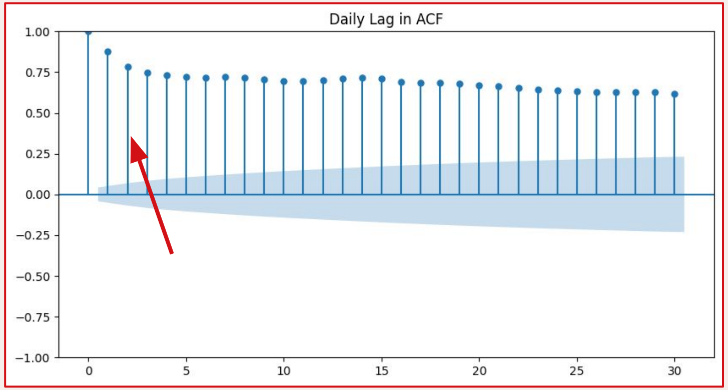


Rolling Statistics

- captures short-term trends

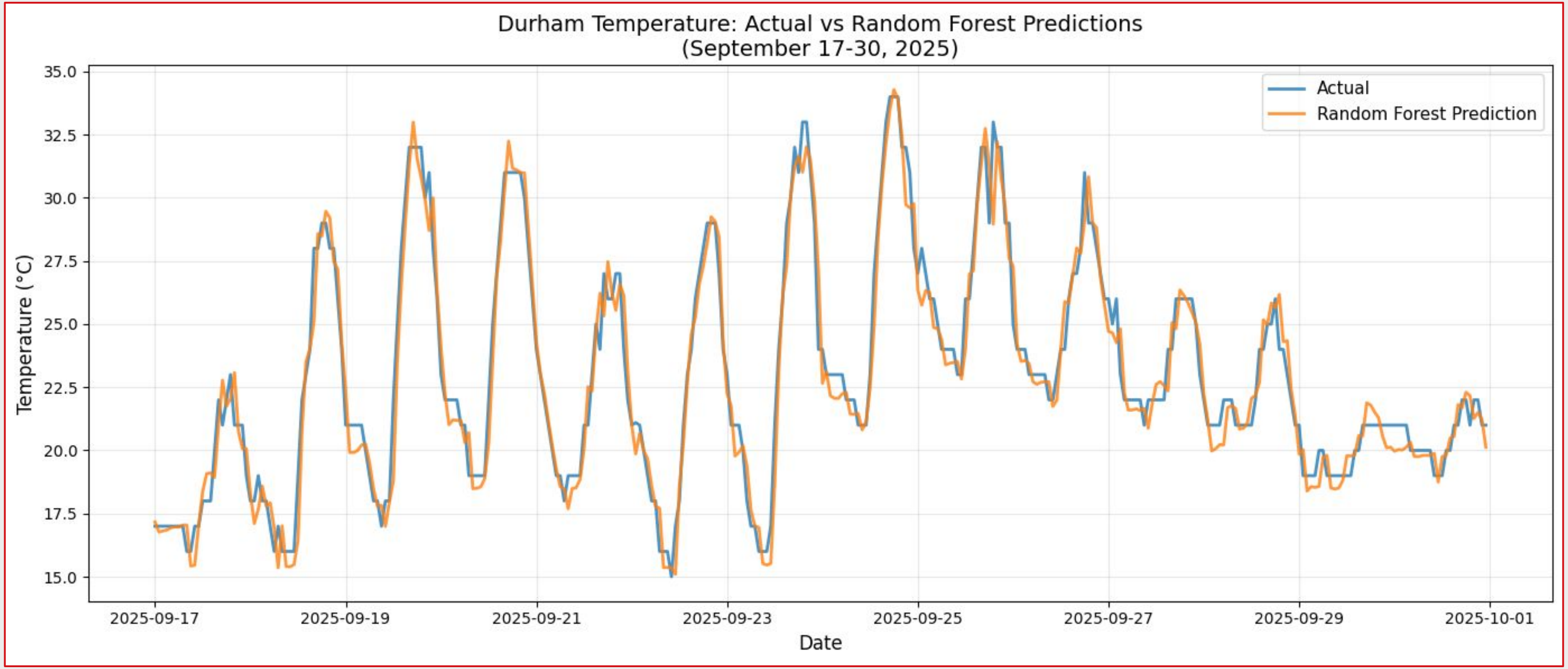
- aggregates temperature over multiple hours instead of one instance

- 24 hour rolling mean and standard deviation
  - weather trend over a day
- 168 hour rolling mean
  - weather trend over the week



Random Forest Model

- Uses multiple decision trees in order to model complex interactions and autoregression
- Hyperparameters: n_estimators = 100, max_depth = 12, min_sample_leaves = 2



Evaluation Approach

How did we evaluate the performance of the models?

1

Checked linear Assumptions

2

Compared with the baseline Linear Regression model

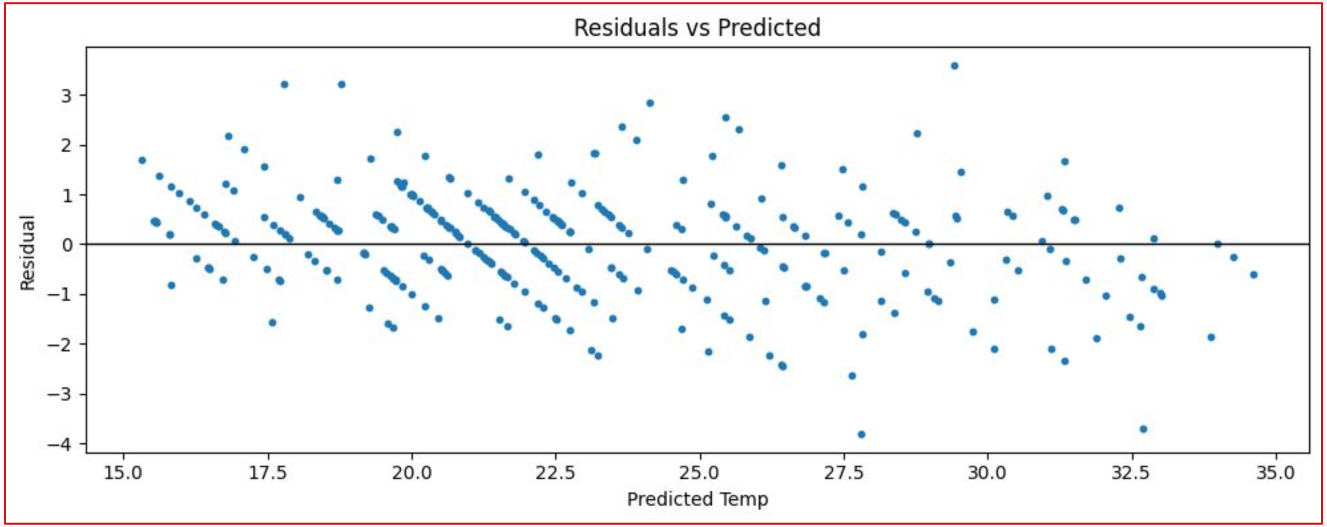
3

Used Time Series Cross Validation to finetune Random Forest hyperparameters

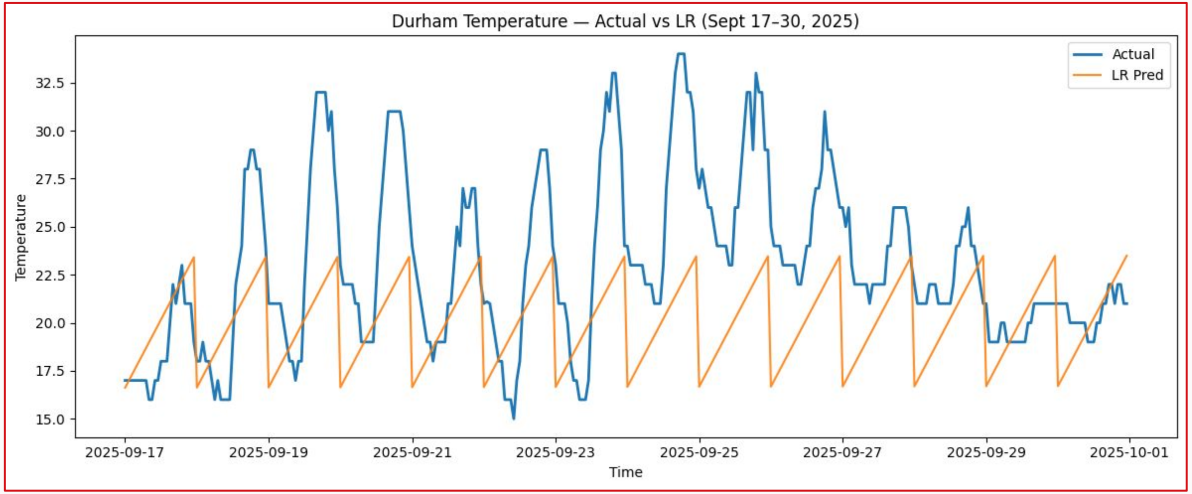
4

Analyzed and compared MSE, MAE, R-squared

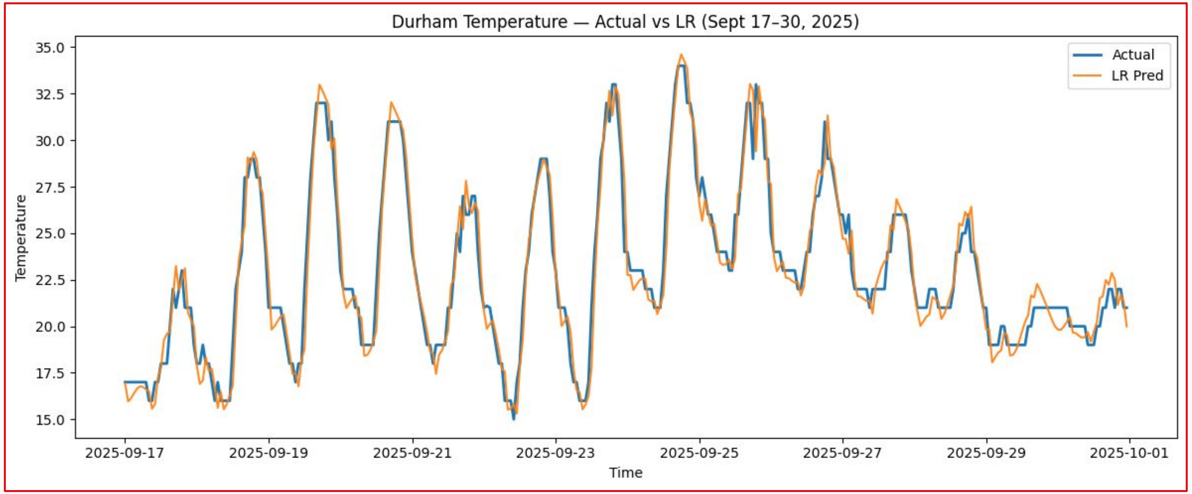
Linear Assumptions



- Linearity: residuals have a pattern
- Homeodascicity: variance of residuals are somewhat equally distributed



Baseline Model



Performance

	Linear Regression	Linear Regression (Kyoto)	Random Forest
MSE	1.11	43.95	0.952
MAE	0.82	5.58	0.756
R^2	0.94	-1.28	0.95

Thank you